Interactive Visualizations to Improve Bayesian Reasoning

Jennifer Tsai[1], Sarah Miller[2], and Alex Kirlik[1]
[1]University of Illinois at Urbana-Champaign, Urbana, IL
[2]The Charles Stark Draper Laboratory, Cambridge, MA

Proper Bayesian reasoning is critical in a variety of domains that require practitioners to make predictions about the probability of events contingent upon earlier actions or events. However, much research on judgment has shown that people who are unfamiliar with Bayes' Theorem often reason quite poorly with conditional probabilities due to various cognitive biases. Owing to previous successes of visualization techniques for debiasing judges and improving judgment performance, we created an interactive computer visualization designed to aid Bayes-naïve people in solving conditional probability problems that would not require a training period to use, and would be flexible enough to accommodate many problem types. Results are suggestive that participants using our interactive visualization were able to substantially improve their Bayesian reasoning performance above that of previous debiasing methods. This finding has significant implications for expanding the toolbox of techniques that can be used to more accurately elicit predictions and forecasts from judges whose expertise lies beyond the realm of statistics.

INTRODUCTION

Bayesian reasoning problems, where one must calculate or otherwise infer the conditional probability of a hypothesis based on how new information or evidence updates its prior probability, are surprisingly common in everyday life. For example, they have recently become front and center in the medical field – specifically, the interpretation of health statistics by doctors and their patients (Gigerenzer et al., 2007). It has been demonstrated many times over that the doctors who administer all sorts of diagnostic tests often have very little understanding of what exactly a positive test result means (Hoffrage & Gigerenzer, 1998). This lack of understanding is then passed on to the patient, who must make potentially life-altering decisions, medical or not, based on an incomplete and often incorrect understanding of the results, and with potentially devastating or fatal outcomes.

Other domains where Bayesian problems are commonly observable include weather forecasting and intelligence analysis. For example in hurricane prediction, the numbers given for the chances of a hurricane striking a particular area of land are often interpreted differently by meteorologists versus laypeople, which can result in massive loss of life when forecasting experts and residents view the same evidence, but disagree about the urgency of evacuation calls. Intelligence analysts are often charged with judging the probability of critical global events that are conditional upon earlier events and actions (Matheny, 2010). For example, an analyst might be asked to estimate the probability that there will be a military coup in some foreign country, given that a particular candidate wins the upcoming presidential election in that country.

Given the vast national and international consequence of such problems that span a wide range of fields, it is clearly of great importance that people are able to make correct/timely Bayesian inferences when called for, which would seem to require possessing a large amount of statistical knowledge. However, potential issues arise in that in many work domains, the judges tasked with solving these problems are exactly the ones who may not necessarily possess a high level of expertise in mathematics or statistics. Even in domains with highly quantitative aspects such as meteorology and hurricane forecasting, these judges may have limited-to-no experience with thinking in terms of priors and base rates, converting their knowledge into numerical probabilities, or aggregating or combining probabilities. Compounding the problem is that explicitly teaching concepts such as Bayes' Theorem is a daunting task, as many a statistics educator has experienced, and there is often little time for students to learn how to properly use and interpret such complex numerical concepts.

Judgment and Debiasing Research

Several decades worth of psychological research has shown that in the context of reasoning with conditional probabilities, people who are unfamiliar with the advanced probability concept of Bayes' Theorem often perform quite poorly, owing to a variety of cognitive biases. In their (1974) research, Tversky and Kahneman describe the phenomenon of base rate neglect, which occurs when an inaccurate judgment is made on the conditional probability of an event given some evidence, due to the judge not having taken into account the prior probability (base rate) of that event. However, it is unquestionable that people offer knowledge that serves as both necessary and valuable input for forecasting scenarios. For instance, human judgment is potentially more flexible and adaptable to situational changes than math models alone, by being able to take into account the newest of information that has yet to be incorporated into the models (Hansell, 2008). The overarching goal then, is to tap into the expertise of people in the judgment process, without introducing any possible negative effects due to their cognitive biases.

Indeed, recent research programs have gone beyond trying to understand cognitive biases, to investigating methods to reduce, eliminate, or mitigate them. Debiasing research (e.g., Arkes, 1991; Larrick, 2004) has put forth techniques such as "consider the opposite" to combat confirmation bias, and

stepwise procedures for eliciting more accurate confidence intervals (Soll & Klayman, 2004). These, and other similar debiasing methods, have been shown to reduce overconfidence and result in better calibrated judgments.

Frequency Formats

To date, one of the most successful debiasing techniques for aiding judges specifically in the context of Bayesian reasoning is the use of frequency formats, whereby the probabilities in a contingent estimation problem are reframed in terms of natural frequencies (Gigerenzer & Hoffrage, 1995). This essentially involves converting a phrase such as, "the probability of being HIV+ is 0.1%," into the analogous phrase, "1 out of every 1000 people is HIV+." While the mathematical underpinnings of the problem remain the same, judges appear to perceive the two framings in very different psychological manner. Ordinarily when Bayes-naïve people attempt to solve Bayesian problems where the provided statistics are framed, as is usual, in terms of probabilities, they can only do so approximately 16% of the time. With frequency formats however, reframing the statistics of the problems in terms of natural frequencies boosts accuracy up to about 46%, in part by discouraging base rate neglect. The means by which it is able to accomplish this seem to be that natural frequencies are computationally simpler, and further, they intrinsically carry information about base rates whereas probabilities do not.

Visualizations

While frequency formats demonstrate an obvious and marked improvement over probability formats, one might still notice and object to the fact that a 46% accuracy rate of elicited judgments is nevertheless strikingly suboptimal, especially when considering the gravity of the events and actions people such as intelligence analysts or hurricane forecasters could be required to make predictions about.

The concept of creating/leveraging external visual aids to support and improve human performance in complex tasks is quite familiar to human factors. In terms of representing aggregated data, visualizations provide a way to potentially communicate not only more than the raw data alone, but to also do so in a way that makes the compiled data simpler, easier to manipulate/read, and sometimes even draws attention to important areas and patterns in the data. Current techniques span a wide range of fields. A number are specifically tailored towards expressing uncertainty in judgment environments, such as Lefevre et al.'s (2005) use of hue, transparency, and opaqueness to represent the height of cloud ceilings, and Finger and Bisantz's (2002) use of degraded icons to model the amount of uncertainty in an enemy threat assessment task.

Specific to the original, single-point estimation Bayesian problems framed in terms of probabilities, Burns (2004) proposed the concept of Bayesian Boxes, which subdivide and use different axis dimensions of a square to represent numerical probabilities for the subset of Bayes problems where the sensitivity and specificity of the test/evidence are precisely equal.

For frequency-formatted Bayesian problems, and without value constraints on sensitivity/specificity, a small number of studies tested whether training participants by showing them how to solve problems represented by one of two different frequency-based, static visualizations, could improve their judgment performance in a subsequent testing session where the participants received probability-formatted problems (e.g. Sedlmeier & Gigerenzer, 2001). These visualizations included a frequency tree diagram, and a frequency grid, a box-based diagram that was essentially squares composed of grids of smaller squares, some of which might be shaded or otherwise marked to denote different values. However, the training manipulation in this study necessitated a 1-2 hour period prior to the testing phase, the visualizations were not provided during testing, and the primary aim of these studies was to improve performance on probability-formatted problems, an unwieldy framing that has already been shown time and again to be detrimental to human Bayesian reasoning.

Similar to past studies of frequency formats, our primary goal was to encourage Bayes-naïve judges to reason in a Bayesian manner at the highest accuracy rate achievable. In contrast however, we propose to go beyond the 46% benchmark set by the frequency format intervention via the introduction of interactive visualization techniques, while also adhering to a number of operating constraints. Owing to the conditions and time limits under which many judges such as intelligence analysts and weather forecasters must operate, any general-use visual aid would have to be easily understood and intuitive enough so that there would be no need for a lengthy training period, for learning either the intricacies of Bayes' Theorem, or understanding any operational or representational aspects of the visualization. In addition, it would need to be flexible enough to accommodate and represent the large space of potential problems that could occur in the world.

EXPERIMENT

We conducted a study to see whether participants who are unfamiliar with Bayes' Theorem would be able to make real-time use of interactive computer visualizations to help them more accurately solve conditional probability problems framed in terms of natural frequencies – the current best format known for inducing Bayesian reasoning in Bayes-naïve judges.

Method

Participants. Thirty-six participants from several neighboring universities and local communities were recruited to participate in this study. Their ages ranged from 18-32 years, and most had completed a college education. They were recruited for a single one hour session and were paid $10 for their time/effort. None of the participants were knowledgeable about conditional probability or Bayes' Theorem.

Design. Participants were tasked with solving six conditional probability problems, adapted from Gigerenzer and Hoffrage (1995), and spanning a variety of surface topics and base rate values. The problems were printed on paper, one to a page, and participants were asked to use a "write-aloud"

protocol while working through each problem. The "write-aloud" protocol essentially consisted of writing down thoughts, calculations, diagrams, or any other tools used to find a solution, in the blank space below each printed problem. The purpose of the protocol was to be able to better track the process by which participants arrived at their answers, and to make sure when participants were actually using algorithms derived from Bayes' Theorem, rather than arriving at correct or near-correct answers by chance/guessing.

The study was a between-subjects design that included 12 participants in each of three conditions. Two of these were the standard probability and frequency formats, where the statistics in the six problems were either framed all in terms of probabilities, or all in terms of natural frequencies. For the third condition, the six problems were framed in terms of natural frequencies, but these participants also received accompanying interactive computer visualizations, one for each problem, to aid them in their judgments.

Interactive visualizations. The visualizations, one per problem plus an example for demonstration purposes, were created in Microsoft Excel 2007 using VBA macros. Each was intended to help judges literally see the relationships between the different components of the Bayesian problem represented. On the surface, they share minor similarities with the static frequency grid training tools in Sedlmeier and Gigerenzer (2001), though ours were created independently and contain interactive properties, as well as are used in very different manner due to differences in study rationale and design. The fundamental basis behind our visualization is a large frequency box diagram, subdivided into many smaller squares that symbolize the entire population or event of interest in the problem. In other words, if a certain population of people is of interest, then each small box within the larger one can be viewed as representing one person in that population.

Figure 1 is the exact example visualization used in this study, for the problem of the number of actual HIV+ patients, given a sample of patients with positive diagnostic test results. It is a 10 × 10 box which represents a group of 100 patients. In addition to the smaller square units, darker lines within the box demarcate the different major components of the problem, which are in turn color-coded to aid in their visual separation. Areas where there is overlap between a represented group (i.e. HIV+ patients) and a given property of interest (i.e. a positive diagnostic test result) are bicolored in a diagonal striped pattern, with one color belonging to the basis group, and the other belonging to the property of interest.

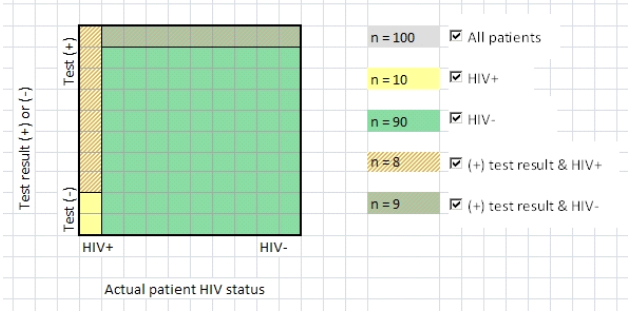

Figure 1. Visualization for the example problem of the number of actual HIV+ patients, given a sample of patients with positive test results.

The most critical component of the visualizations is their capacity for interactivity. Rather than being just a set of passive, completed pictures with accompanying legends, a set of checkboxes to the right of each diagram instead allows for active toggling on/off of the components of a problem in any order or combination.

At start, all checkboxes except the one representing the entire group are unchecked. The diagram is grayed out (Fig.2).

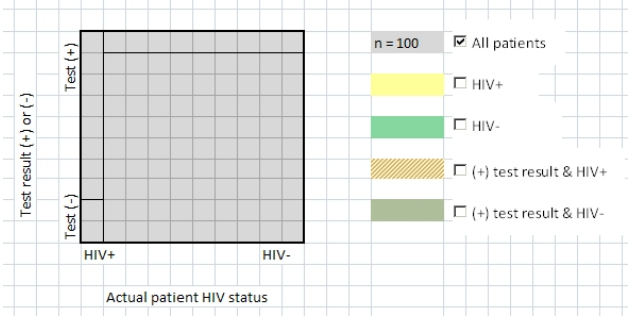

Figure 2. Visualization for the example with all checkboxes unchecked.

The diagram can be incrementally colored in by checking each checkbox that corresponds to a section of the diagram. For example, checking the "HIV+" checkbox in Figure 2 will highlight the 10 patients who have HIV out of the entire sample of 100 patients. Checking the "(+) test result & HIV+" checkbox will highlight the 8 patients who have HIV and also received a positive test result, out of the sample of 10 patients who have HIV. Checking these two checkboxes, along with the "HIV-" one, would result in Figure 3.

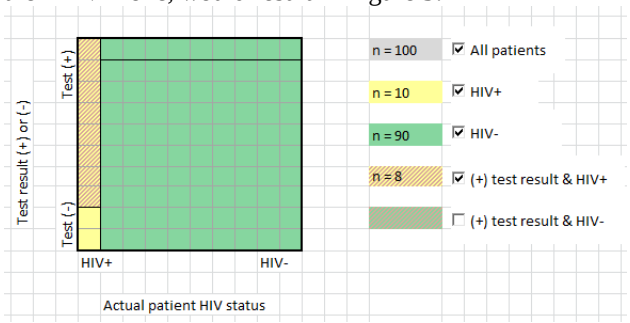

Figure 3. Visualization for the example with first three checkboxes checked.

Though the initial tendency might be to go straight down the line in checking the checkboxes, the boxes can actually be checked and unchecked in any order or combination, a fact that participants were made aware of. For example, checking

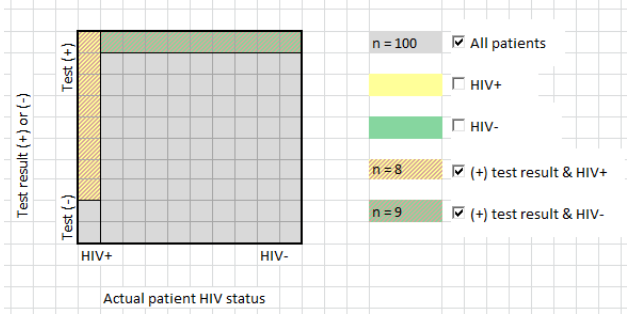

Figure 4. Visualization for the example with first two checkboxes unchecked, and last two checked.

only the "(+) test result & HIV+" and "(+) test result & HIV-" checkboxes would result in Figure 4.

This type of visualization is relatively simple – since the diagram is basically only visualizing the different components of a conditional probability problem, this simplicity allows it to be adapted to a wide variety of problems and topics, and additionally should shorten the time needed for first-time users to understand what the parts of it mean and how they work.

Procedure. Immediately prior to beginning the study, participants in all three conditions were briefed on the example problem, which they were not required to solve, and were not provided the answer to. The text of this problem is reprinted below in both probability and frequency/visualization format.

Probability. "A person in a moderately high-risk population for HIV visits his doctor. The doctor informs him that the probability of a patient like him having HIV is 10%. The doctor also tells him that if a patient has HIV, the probability that an HIV test will correctly identify him as being HIV positive is 80%. However, if a patient does not have HIV, the probability that the test will incorrectly identify him as being HIV positive is 10%. Imagine a patient with a similar lifestyle to this person. This patient takes the HIV test, and the result comes back positive. What is the probability that the patient is actually infected with HIV?"

Frequency/visualization. "A person in a moderately high-risk population for HIV visits his doctor. The doctor informs this person that the probability of him having HIV is about 10 out of every 100 patients. The doctor also tells him that out of those 10 patients with HIV, an HIV test will correctly identify 8 of them as being HIV positive. However, out of the 90 patients who do not have HIV, the test will also incorrectly identify 9 of them as being HIV positive. Imagine a group of patients with similar lifestyles to this person. They take the HIV test, and all of their results come back positive. How many of these patients are actually infected with HIV?"

Participants in the visualization condition additionally received the accompanying visualization for this example problem (Figure 1) and a short explanation lasting 1-2 minutes on how to use the interactive aspects, as well as the meaning behind the different parts of the diagram. Since the visualizations for each of the six test problems were of the same general format as the example visualization, no further instruction was provided for the test visualizations.

Participants were told to work the problems in the order in which they appeared in their paper packets. Ordering of problems was randomized. Participants were also told that they would not be timed, and that they should try to solve the problems to the best of their ability.

RESULTS

Participants' write-aloud protocols were examined to confirm the presence/absence of normative Bayesian reasoning processes. If the protocol for an answer did not adhere to some form of proper Bayesian reasoning, then the answer was counted as incorrect. This criterion disallows final answers that are close to being correct merely by chance, while allowing answers with some measure of calculation or rounding error (provided that a Bayesian algorithm was demonstrated). Note that this answer correctness judgment is purely objective. Test questions have explicitly defined probabilities or frequencies; thus, solutions are well-defined from the equation of Bayes' Theorem – for each problem, there is only one correct answer, and one mathematically correct way to reach it (there may be several mathematically equivalent ways, but only one general way, such as in the sense that $2+2=4 \approx 4/2+4/2=4$). A protocol constitutes evidence of Bayesian reasoning only if it contains the exact math necessary to normatively solve the problem.

Results of this study show that participants using the frequency format demonstrated a higher proportion of correct Bayesian judgments than those using the probability format, and participants using the visualization format demonstrated an even higher proportion of correct judgments on top of those using the frequency format: 75% for visualization, 54% for frequency, and 31% for probability. Due to the normality assumption not being sound for this data, non-parametric tests are preferable. A Kruskal-Wallis test yielded significant differences among the three conditions ($p < 0.05$). For pairwise comparisons, a Mann-Whitney test yielded a significant difference between the probability and visualization conditions ($p < 0.05$), and marginally significant differences between the probability and frequency conditions ($p = 0.12$), and the frequency and visualization conditions ($p = 0.08$).

Though the two intermediate pairwise comparisons were only marginally significant, this was likely due to our sample size being several times smaller than those used in previous frequency format studies. To date, no one questions the repeatedly proven effectiveness of frequency formats over probability formats. The evidence for a corrective effect of frequency formats over probability formats, and a further corrective effect of visualization formats over frequency formats, becomes even stronger upon viewing Figure 5, which shows percentage of accurate Bayesian judgments, broken down by individual test problem. The six problems in this study were specifically chosen to span a large range of base rates, from 0.05% (Wall St. Exec) to 36% (School Admission). From this figure, it can be seen that the directional data trend by condition (probability, frequency, visualization) is perfectly consistent across all six problems, with probability being the worst, frequency being better, and visualization being the best.

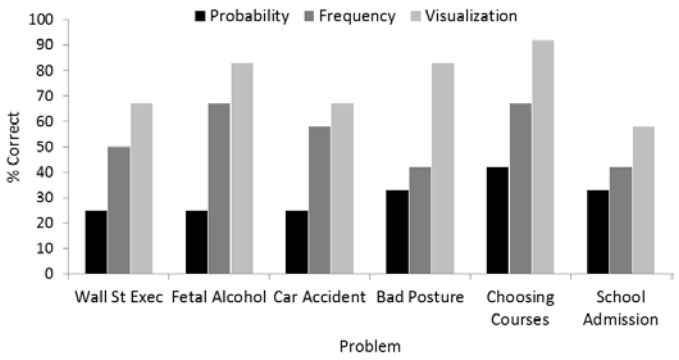

Figure 5. Percent accuracy of Bayesian reasoning, broken down by individual test problem. The problems are ordered, left to right, from smallest (0.05%) to largest (36%) base rate.

DISCUSSION

In summary, the results of our study exactly mirror the predicted ordering of probability being worst, frequency in the middle, and visualization being best, for every problem tested. Perhaps surprisingly, the differences between probability and frequency, and between frequency and visualization, failed to reach significance. The consistency of the directional trends remains encouraging however, and most likely points not to a failure of either the frequency or visualization interventions, but rather a need to follow up by reinvestigating matters with a larger sample size – one more akin to those used in previous frequency format studies – for more statistical power.

Why might the use of simple interactive visualizations so consistently and so readily boost judges' Bayesian reasoning accuracy? The particular visualizations used in this study did not contain any information additional to or beyond what was provided in the problem texts of the probability and frequency format conditions. All numbers appearing beside checkboxes in a visualization are directly readable from the frequency version problem text, and appear as probabilities in the probability version problem text. However, actually being able to see these numbers in the visualizations may provide visual reinforcement of verbal/ textual understandings of their relations, or perhaps even help people to understand problem components they might not have comprehended otherwise from text alone. When one observes the two shaded sections of Figure 4, it is visibly apparent from the comparative areas of the two sections that the number of patients with a positive test result who are actually infected with HIV is roughly equal to the number of patients with a positive test result who are not actually infected with HIV. The observation of this relation effectively eliminates incorrect intuitions that a test having relatively high sensitivity and specificity must also have a high degree of accuracy. We argue that providing opportunity for judges to observe these visibly apparent relations is precisely the benefit of visualizations. Moreover, the beneficial effect seems impressively robust, with visualizations apparently able to improve reasoning performance even under conditions where judges receive minimal training on using them.

CONCLUSION

As has been found previously, Bayesian reasoning can be encouraged by presenting information in a natural frequency format as opposed to a probability format. The current study newly uncovers evidence that this improvement can be further bolstered via the introduction of interactive computer visualizations to help judges literally see and manipulate the different parts of a Bayesian probability problem. In addition, it is not necessary to spend large amounts of time teaching judges to use these visualizations prior to eliciting their contingent judgments. Simply supplying a visualization and short explanation of its parts, without ever showing the judge how to use it to solve a problem, can be sufficient. All of this, coupled with the simplicity and flexibility of the visualizations used in this study, make a strong case for considering interactive computer visualizations as a formidable debiasing technique, to be added to the toolbox of techniques that can be used to more accurately elicit predictions and forecasts from judges whose expertise may lie beyond the realm of statistics.

The present work was a novel attempt to use visualization techniques, on top of other previously proven methods, to maximize the accuracy of Bayesian reasoning exhibited by Bayes-naïve judges under conditions of minimal training. While just one particular visualization was tested here, the possibilities for variations, or even completely different types of visualizations, await investigation. Future work could focus on tweaking aspects of the visualization created for this study, or perhaps on testing alternate forms of visualizations such as interactive frequency trees or water-pipe diagrams. Another potential direction of interest might be to see if the concept of aiding via visualization could be extended from the elementary single-point judgments tested here, to Bayesian problems of more complicated form, such as ones involving chains of reasoning or multiple pieces of evidence.

ACKNOWLEDGMENTS

This research was supported by a grant from the Charles Stark Draper Laboratory to the University of Illinois.

REFERENCES

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486-498.

Burns, K. (2004). Bayesian Boxes: A colored calculator for picturing posteriors. *Proceedings of the 3rd International Conference on the Theory and Application of Diagrams.*

Finger, R. A., & Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science, 3,*1-25.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8,* 53-96.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Hansell, S. (2008, September 18, 2008). How wall street lied to its computers. *New York Times.*

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73,* 538-540.

Larrick, R. P. (2004). Debiasing. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316-337). Oxford, England: Blackwell.

Lefevre, R. J., Pfautz, J., & Jones, K. (2005). *Weather forecast uncertainty management and display.* Paper presented at the 21st International Conference on Interactive Information Processing Systems (UPS) for Meteorology, Oceanography, and Hydrology, San Diego, CA.

Matheny, J. G. (2010, June 30). Aggregative Contingent Estimation (ACE) program - Broad Agency Announcement (BAA). Retrieved from http://www.iarpa.gov/solicitations_ace.html.

Sedlmeier, P. & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130,* 380-400.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 229-314.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.